

University of Groningen

Comparative genomics of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* using a *Streptomyces coelicolor* microarray system
Hsiao, Nai-hua; Kirby, Ralph

Published in:

Antonie Van Leeuwenhoek: International Journal of General and Molecular Microbiology

DOI:

[10.1007/s10482-007-9175-1](https://doi.org/10.1007/s10482-007-9175-1)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2008

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Hsiao, N., & Kirby, R. (2008). Comparative genomics of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* using a *Streptomyces coelicolor* microarray system. *Antonie Van Leeuwenhoek: International Journal of General and Molecular Microbiology*, 93(1-2), 1-25. <https://doi.org/10.1007/s10482-007-9175-1>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Comparative genomics of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* using a *Streptomyces coelicolor* microarray system

Nai-hua Hsiao · Ralph Kirby

Received: 9 April 2007 / Accepted: 11 May 2007 / Published online: 21 June 2007
© Springer Science+Business Media B.V. 2007

Abstract DNA/DNA microarray hybridization was used to compare the genome content of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* with that of *Streptomyces coelicolor* A3(2). The array data showed an about 93% agreement with the genome sequence data available for *S. avermitilis* and also showed a number of trends in the genome structure for *Streptomyces* and closely related *Kitasatospora*. A core central region was well conserved, which might be predicted from previous research and this was linked to a low degree of gene conservation in the terminal regions of the linear chromosome across all four species. Between these regions there are two areas of intermediate gene conservation by microarray analysis where gene synteny is still detectable in *S. avermitilis*. Nonetheless, a range of conserved genes could be identified within the terminal regions. Variation in the genes involved in differentiation,

transcription, DNA replication, etc. provides interesting insights into which genes in these categories are generally conserved and which are not. The results also provide target priorities for possible gene knockouts in a group of bacteria with a very large numbers of genes with unknown functions compared to most bacterial species.

Keywords Actinomycetales · Genomotyping · Linear genome

Introduction

Streptomyces are a group of aerobic high %G+C Gram positive bacteria that undergo complex differentiation to form filamentous mycelium, aerial hyphae and spores. In addition, they produce a broad range of secondary metabolites including antibiotics, antiparasitic agents, herbicides, anti-cancer drugs and various enzymes of industrial importance. Two *Streptomyces* species have had their complete genome sequences published, namely the model organism *Streptomyces coelicolor* (%G+C = 72.1) and avermectin producer *Streptomyces avermitilis* (%G+C = 70.7) (Bentley et al. 2002; Ikeda et al. 2003). Two important aspects of the genomes structures of *Streptomyces* were supported by sequence data. Firstly, that the genome size of *Streptomyces* is large compared to other bacteria; 8,667,507 basepairs for *S. coelicolor* (7,825 protein coding genes) and

Electronic Supplementary Material The online version of this article (doi:10.1007/s10482-007-9175-1) contains supplementary material, which is available to authorized users.

R. Kirby (✉)
Department of Life Science, National Yang-Ming
University, Beitou, Taipei 114, Taiwan
e-mail: rkirby@ym.edu.tw

N.-h. Hsiao
Microbial Physiology, University of Groningen, Kerklaan
30, NL-9751 NN Haren, Groningen, The Netherlands

9,025,608 bp (7,577 protein coding genes) for *S. avermitilis*. Secondly, that the genomes of these two species are linear and both ends contain unique terminal inverted repeats that probably covalently bind a terminal protein. Terminal inverted repeats and covalently bound terminal proteins are not found in the limited number of other bacteria that have linear chromosomes such as *Borrelia burgdorferi* and *Agrobacterium tumefaciens* and, up to the present, seem to be unique to the *Streptomyces* and perhaps other Actinobacteria (Lin et al. 1993; Chen et al. 2002; Goodner et al. 1999; Huang et al. 2004). Over 2,500 *Streptomyces* strains are present in the Ribosomal Database Project (<http://www.rdp.cme.msu.edu>), over 1,500 are available at the American Type Culture Collection (<http://www.atcc.org/>) and many more are held in both public and private culture collections throughout the world. Analysis of the small subunit ribosomal RNA gene sequences of *Streptomyces* confirms that they form a monophyletic clade, but one with considerable diversity. In addition, there is significant gene diversity at the interspecies level across the genomes of both completely sequenced *Streptomyces* with 2,291 gene unique to *S. avermitilis* and 2,307 genes unique to *S. coelicolor*. This makes them particularly interesting targets for comparative genomic studies. In this study we chose four species to begin an analysis of the genomic diversity of the *Streptomyces*. *S. avermitilis* was chosen because of the availability of the complete genome sequence of this species, while *Streptomyces maritimus* was chosen because of its intermediate position in terms of phylogeny within the *Streptomyces*. *Streptomyces cattleya* was chosen because, based on small subunit ribosomal RNA sequence, this species is phylogenetically quite divergent from *S. coelicolor* and branches near the root of the *Streptomyces* clade. *Streptomyces cattleya* is a β -lactam producing species. Finally, *Kitasatospora aureofaciens* was chosen as this genus is very closely related to the *Streptomyces*.

The availability of two microarrays for *S. coelicolor* (Lum et al. 2004; Huang et al. 2001; Vinciotti et al. 2005; <http://www.surrey.ac.uk/SBMS/Fgenomics/Microarrays/index.html>) makes possible a comparative genomic analysis of *Streptomyces* species. The genes that make up the genome of *S. coelicolor* have been classified based on scheme of Riley and colleagues for *E. coli* and modified for *S. coelicolor*

(http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml). A microarray analysis of the genomes of these *Streptomyces* using the *S. coelicolor* microarray is able to provide a wide ranging comparative analysis of the conserved genome content of these *Streptomyces*. This type of approach, where a heterologous microarray is used to analyze the genome content of a range of strains or species, has been successfully used in a wide range of organisms (Akman and Aksoy 2001; Akman et al. 2001; Behr et al. 1999; Chan et al. 2003; Cho and Tiedje 2001; Dorrell et al. 2001; Dziejman et al. 2002; Fitzgerald et al. 2001; Gill et al. 2002; Leonard et al. 2003; Murray et al. 2001; Porwollik et al. 2002; Salama et al. 2000; Israel et al. 2001; Rajashekara et al. 2004). The strains analyzed using this approach range from intraspecies comparisons such as *Campylobacter jejuni*, *Vibrio cholerae* and *Staphylococcus aureus* (Dorrell et al. 2001; Dziejman et al. 2002; Fitzgerald et al. 2001) to interspecies comparisons such as *Sodalis glossinidius* versus an *Escherichia coli* array, *Salmonella bongori* versus a *Salmonella enterica* array, *Shewanella* species versus *Shewanella oneidensis* and *E. coli* arrays and *Brucella* species versus a *Brucella melitensis* array (Akman et al. 2001; Chan et al. 2003; Murray et al. 2001; Rajashekara et al. 2004).

In this study, we used both versions of the *S. coelicolor* genome microarrays to compare the gene complements of the three *Streptomyces* species and one *Kitasatospora* species. The genus *Kitasatospora* is closely related to the genus *Streptomyces* in terms of morphology, chemical taxonomy and small subunit ribosomal RNA sequence analysis. Thus, the choice of a species from this genus acts as potential outgroup in terms of overall genome structure. In terms of genes that are conserved, the types of genes of particular interest include genes involved in secondary metabolism, genes involved in chromosome replication, genes in the terminal regions of the chromosome, sigma factors, genes involved in differentiation and hypothetical genes. In terms of gene absence, the distribution of such genes along the chromosome and the apparent absence of any major housekeeping genes in a specific species are of interest. This information provides insights into genes that make up the core complement for a member of the *Streptomyces* and into which genes are central to defining a *Streptomyces* species

Materials and methods

16S phylogeny

This was carried out on selected small subunit 16S ribosomal RNA gene sequences obtained from Ribosomal Database Project-II Release 9 (<http://www.rdp.cme.msu.edu/index.jsp>) and aligned using CLUSTALX (Thompson et al. 1997). The analysis was carried out using Neighbor-Joining algorithm from the same program. In the case of *S. maritimus*, the taxonomy of the strain was confirmed by DNA sequencing of the 16S ribosomal RNA gene.

Arrays

Two series of arrays that cover about 97% of the complete genome of *Streptomyces coelicolor* A3(2) (Lum et al. 2004; <http://www.surrey.ac.uk/SBMS/Fgenomics/Microarrays/index.html>) were used in this study. Both arrays are PCR arrays, but from different sources, namely Stanford University, USA and the University of Surrey, UK and made up of different PCR products. The Stanford array as used in this study contained sequences covering 7603 open reading frames. The Surrey microarray is made up of 7,758 unique PCR amplified sequences, 7,563 from the chromosome and 195 from SCP1. There are an additional 376 non-unique, alternative and cross-hybridizing sequences that are also spotted on to the array together with no probe spots and control spots. The two types of arrays were used to improve validation with a system using heterologous hybridization; however, only the University of Surrey array was hybridized and analyzed in duplicate. The major difference between the two arrays was that the Surrey array did not include a number of transposition element related genes, although there were other overlap differences. The sequences of the PCR products are not available for either array due to intellectual property protection requirements.

Strains and growth conditions

S. coelicolor A3(2) (SCP1⁺) 104, *S. avermitilis* ATCC 31267, *S. cattleya* ATCC 35852, *S. maritimus* Yang-Ming and *K. aureofaciens* ATCC 10762 were used in these studies. Fresh spores were collected and

mycelium cultured in TSB liquid medium with 0.5% glycine at 30°C overnight.

Preparation of labeled DNA

Genomic DNA from a stationary phase culture was purified by the salting out procedure (Pospiech and Neumann, 1995) and had been sonicated to < 2 Kb. Four to six micrograms of sonicated genomic DNA were used as template and this was denatured in the presence of 12 µg of 72%-GC-content random hexamers in a total volume of 25 µl at 100°C for 10 min. The mixture was then snap-cooled on ice before adding the remaining reaction components: 1.5 µl of Cy3-dCTP or Cy5-dCTP (Amersham Pharmacia Biotech), 4µl Klenow fragment (NEB #212), 5µl Klenow buffer, 0.5 µl dNTP (4 mM dATP, 4 mM dTTP, 10 mM dGTP, and 0.2 mM dCTP), and 14 µl ddH₂O. The random primed labeling reaction was carried out for 2–3 h at 37°C. Buffer exchange, purification and concentration of the DNA products was accomplished by three cycles of diluting the reaction mixture in 0.5 ml TE buffer (10 mM Tris and 1 mM EDTA pH 8.0) and filtering through a Microcon-30 microconcentrators (Millipore).

Microarray hybridization and data analysis

The two DNA pools to be compared were mixed and applied to an array in a hybridization mixture that contained 3.68 × SSC, 0.18% SDS, and 1 µg yeast tRNA (total 16.3 µl), which had been heated at 100°C for 5 min before being applied to array. Hybridization took place under a glass coverslip sealed by glue in a humidified Omnislid (Thermo Hybaid) at 60°C for 12–14 h. The slides were washed, dried and scanned for fluorescence using a GenePix™ 4000B scanner (Axon instruments). Average signal intensity and local background measurements were obtained for each spot on each array using GenePixPro software. The dataset was screened for aberrant spots and these were eliminated from the analysis after manual checking. Most genes are present in duplicate on the two arrays and the signal from each pair of spots was inputted into the computer program available from ScanAlyze (Eisen et al. 1998; Gollub et al. 2003).

The data was then processed into a mean \log_2 Cy3/Cy5 ratio format. The dataset was normalized for each array separately and outputted to Excel where after checking the alignment of the datasets from each array, a mean signal for each common gene was calculated. Genes that were absent from either array, mostly transposon related genes in the University of Surrey array, were not included in the analysis. Based on Bentley et al. 2002, the mean signal and standard deviation for the core region of genes from SCO2050 to SCO5800 was calculated. The standard deviation was used to set a cut-off for gene absence at 2SD below the core mean. The microarray data is presented relative to the *S. coelicolor* standard in two ways. This is either as a color plot of the genes where green presents a negative hybridization signal, black represents an equal hybridization signal and red indicates a positive hybridization signal using the program Treeview (Eisen et al. 1998) or as numeric values for the signal from each gene. The microarray data for the four species described here and additional unpublished species can be accessed via rkirby@ym.edu.tw.

Comparison of the microarray dataset for *S. avermitilis* with the complete genome sequence

The nucleotide sequences for all the identified open reading frame from the *S. avermitilis* genome sequence (Ikeda et al. 2003) were compared with the genome sequence of *S. coelicolor* using blastn limiting the output to the best match. This E value dataset for the genes was then aligned with the *S. avermitilis* microarray dataset and a comparison plotted as a scatterplot. Genes showing disagreement between the two datasets were identified based on a 2 Standard Deviation (SD) cutoff for the microarray dataset and a E-10 cutoff for the blast value.

Analysis of gene presence across the chromosome

A graphical display was created by counting the number of gene detected as present from the signal based on the 2SD cutoff from each normalized microarray dataset using a moving window of 10 genes in steps of one.

Results and discussion

Comparison of *S. avermitilis*, *S. cattleya*, *S. maritimus* and *K. aureofaciens* with the *S. coelicolor* genome

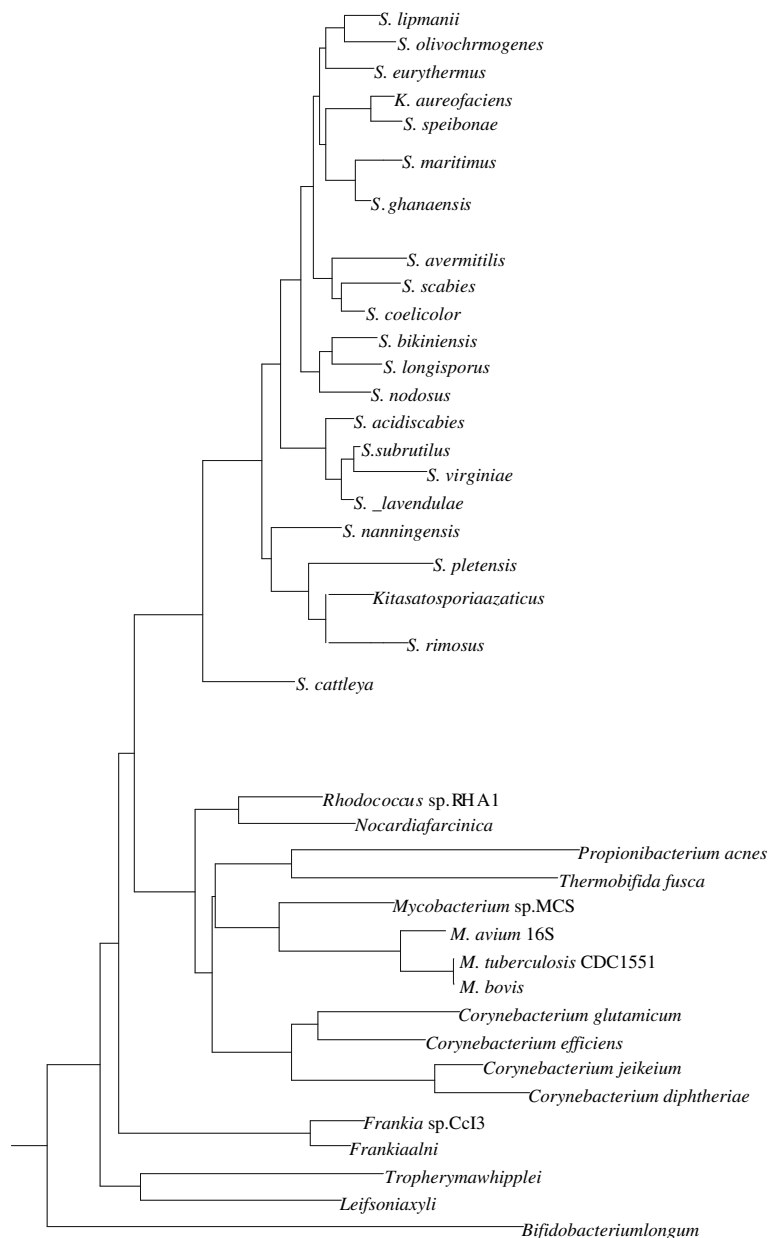
In total, after spot and data validation, a total of 7,083 open reading frames were included in this analysis as presence on both types of array and giving analyzable signal on all three arrays. Validity in this study was initially obtained by using microarrays from two sources that presumably use different PCR products to create the arrays. In addition, the University of Surrey array was hybridized and analyzed in duplicate. In terms of gene absence based on two standard deviations as described in the “Materials and methods” section, the agreement between the Stanford array and the duplicated University of Surrey array was about 95%, while the agreement between the two University of Surrey arrays was about 98%. In order to minimize the effect of divergent individual array spots, the signal mean for each gene from the three arrays was used throughout this study.

In this study, the genomic content of three *Streptomyces* species and one *Kitasatospora* species with divergent taxonomy, antibiotic production and SSU rRNA sequence are compared using two different *S. coelicolor* microarrays. It is clear that there are inherent limitations to this approach. Firstly, only gene absence or divergence rather than the presence of new genes can be identified. Secondly, it is not possible to clearly separate the absence of a gene from the presence of a divergent homologue of the same gene. Finally, although the order of the genes in *S. coelicolor* and *S. avermitilis* are known from their complete genome sequences and are well conserved, this does not mean that the synteny of most of them is conserved in other *Streptomyces* species. However, the detection of synteny across Actinobacteria including *Mycobacterium tuberculosis*, *Corynebacterium glutamicum* and other species (Bentley et al. 2002 and unpublished data) supports a conserved central core structure to the genomes of the Actinomycetes and a priori most *Streptomyces*. Thus, although major chromosomal reorganizations in the central core region cannot be detected by microarray data, a basic chromosomal structure can be assumed as a first approximation; namely, a linear chromosome with variable terminal regions and a relatively well conserved core region.

When the pooled data from the two arrays for the four species was analyzed using Cy-3 labeled *S. coelicolor* A(3)2 chromosomal DNA compared to heterologous Cy-5 labeled chromosomal DNA, a wide range of signal variation could be noted and this is shown in Supplementary Fig. 1. The SSU rRNA tree places the divergence of these four strains from *S. coelicolor* as *S. cattleya* > *K. aureofaciens* > *S. maritimus* > *S. avermitilis*

(Fig. 1). Gene differences were present in the order *S. cattleya* > *K. aureofaciens* > *S. avermitilis* > *S. maritimus* based on $-2SD$ cutoff below the mean signal for the core region genes. The microarray data thus shows general agreement with *S. cattleya* and *K. aureofaciens* being more divergent and the other two species being relatively closer. It is interesting to note that the *Kitasatosporia* species used in this study, *K. aureofaciens*,

Fig. 1 SSU rRNA phylogenetic tree of selected *Streptomyces* species and other Actinomycetes that have known complete genome sequences. The species analyzed by microarray are indicated in bold



shows the same general structure as the *Streptomyces* species. This is not unexpected and confirms the close relationship between *Kitasatospora* and *Streptomyces* and agrees with the SSU rRNA tree data.

Further support of the reliability of the data comes from a comparison of the blastn E values for all genes and the microarray data as shown in the Fig. 2 scatterplot. This indicated 232 out of 6,832 genes show gene absence by microarray when they seem to be present by blastn and 268 out of 6,832 gene show gene presence by microarray when they seem to be absent by blastn; these results are both based on cutoffs of $-2SD$ for the microarray data and -10 for the E value. This gives an overall reliability for *S. coelicolor* compared to *S. avermitilis* of 93%. Potential errors factors include in the case of the former type of error, poor spotting of the array at that point and choice of the PCR product sequence (the comparison is with the whole gene, as the PCR products are not available) and in the latter case cross-hybridization between multiple gene copies or a unreliable hybridization signal due to poor washing in that area. However, the results for *S. avermitilis* clearly support the reliability of the genome comparisons produced by this study.

Distribution of gene differences across the complete chromosome of *S. coelicolor* for all four other *Streptomyces* species

The whole chromosome microarray dataset supports the following structure for the *Streptomyces* chromosome. Based on Fig. 3 and Supplementary Fig. 1, there is a central core of conserved probably syntenous genes that can be found across many Actinomycetes and in the *S. coelicolor* genome this reach from about SCO2050 to SCO5800 (Bentley et al. 2002). The regions between SCO1100 and SCO2050 and between SCO5800 and SCO7600 are also quite well conserved between the *Streptomyces* studied here as well as being syntenous between the *S. coelicolor* and *S. avermitilis* genome sequences. However they are not present when the genomes of these two species are compared bioinformatically to other divergent Actinomycetes. These two regions seem to be two genus specific areas. Figure 3 also clearly shows that gene conservation drops off dramatically in the terminal region. The regions from the left terminus to SCO1100 and from SCO7600 to the right terminus show much higher gene divergence that the rest of the chromosome. This agrees with the results for the *S. ambofaciens* sequencing studies of that species' terminal regions (Choulet et al. 2006a, b).

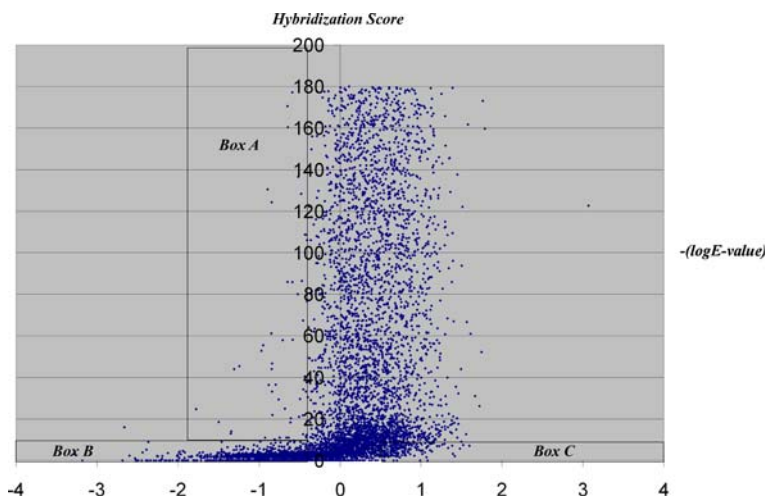


Fig. 2 Scatterplot comparing gene presence/absence based on the microarray data and gene presence/absence based in blastn between *Streptomyces coelicolor* and *Streptomyces avermitilis*. See “Material and methods” for details. Box A and Box C includes genes identified as absent in *S. avermitilis* by the

microarray dataset but present using blastn and genes present in *S. avermitilis* using blastn, but identified as absent by the microarray dataset

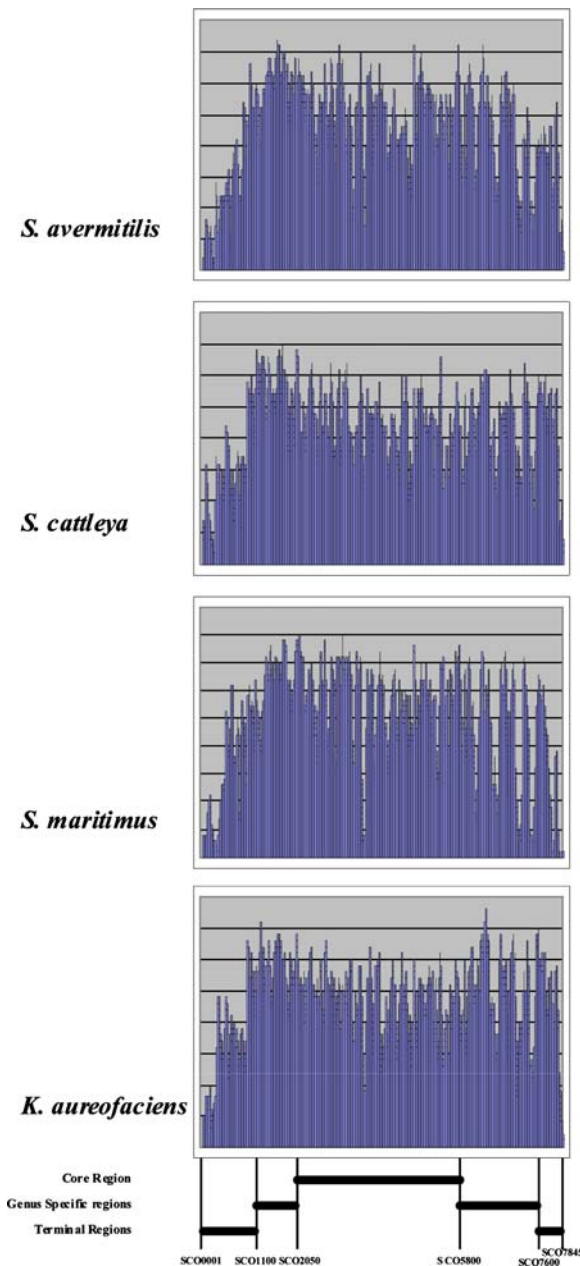


Fig. 3 Analysis of “gene presence” across the four species. Created using a moving window of 10 genes and counting the number of genes with a microarray signal $>2SD$ below the mean for the core region genes. The Y axis is the count for “gene presence”

The gene conservation levels averaged across the four species are as follows: left terminal region (SCO0001–SCO1100) 40.9%; left genus specific region (SCO1101–SCO2050) 84.8%; core region (SCO2050–SCO5800) 79.4%; right genus specific

region (SCO5801–SCO7600) 69.6% and right terminal region (SCO7601–SCO7845) 50.3%. It is noticeable that neither the size nor the distribution of conserved genes is symmetrical between the two terminal regions or the two genus specific regions. Notably, the genus specific region actually has a higher frequency of gene conservation than the core regions as a whole and that the left terminal region is much larger than the right terminal region. This possibly represents horizontal exchange of terminal regions by recombination between strains/species that involves only one terminal region. Such an event would give rise to asymmetric gene conservation similar to that detected here.

In the Karoonuthaisiri et al. (2005) study of regional gene expression in *S. coelicolor*, the boundaries for higher transcript levels during vegetative growth were placed at 1.5 Mb for the left arm and 2.3 Mb for the right arm. The former is midway across the left genus specific region and the latter approximately agrees with the boundary between the core and the right genus specific region. As the core region boundaries are also defined in terms of synteny with the *Mycobacterium* and *Corynebacterium* genomes as well as the data presented here, this supports the idea that the *S. coelicolor* chromosome structure is asymmetrical with respect to both gene conservation and gene function. It should be noted that because we are using only *S. coelicolor* as the source of the array data, the results do not imply that the genomes of *S. cattleya*, *S. maritimus* and *K. aureofaciens* are asymmetric. However, it should be noted that the *S. avermitilis* genome is also asymmetric (Ikeda et al. 2003).

Notably, there are 22 identifiable regions where all four species show a significant degree of concurrent gene absence outside of the terminal regions (Table 1). The regions of high gene divergence are shown in Supplementary Fig. 2 in detail. Previously, Bentley et al. identified 14 regions in the *S. coelicolor* chromosome that were potentially laterally acquired regions. This analysis pinpoints all of these regions and quite accurately, usually to within one or two open reading frames. This suggests that other eight regions are probably quite robust when designated as potential lateral transfer regions. It also supports the usefulness of the microarray approach. All 22 regions were analyzed using Frame Plot (Artemis v7.1) and except for region B, they show abnormalities for at

Table 1 Areas of the *Streptomyces coelicolor* genome identified as potentially horizontally transferred regions based on microarray parallel gene absence in all four species

Region	Area of chromosome	Genes missing ^a	% ^a	Significant features
Region A	SCO0996–SCO1010	17/29	59	Integrase, insertion sequence
Region B	SCO2860–SCO2879	53/76	69	Rifampin ribosyl transferase
Region C	SCO3249–SCO3288	94/156	60	Integrase, excisionase
Region D	SCO3471–SCO3538	198/268	73	Agarase
Region E	SCO3584–SCO3599	30/60	50	
Region F	SCO3929–SCO3937	22/32	68	Integrase/recombinase, <i>fstK</i> -like
Region G	SCO3980–SCO4001	56/64	88	Hypothetical proteins
Region H	SCO4052–SCO4066	132/144	92	Boundary <i>dnaZ</i> gene
Region I	SCO4210–SCO4223	37/54	69	
Region J	SCO4247–SCO4257	21/36	58	Hypothetical proteins
Region K	SCO4340–SCO4354	34/40	85	Integrase, DNA invertase
Region L	SCO4509–SCO4547	106/144	74	Hypothetical proteins
Region M	SCO4613–SCO4631	40/68	59	Integrase, excisionase
Region N	SCO4686–SCO4700	24/44	55	Boundary ribosomal proteins operon
Region O	SCO5323–SCO5351	57/80	71	Integrase, excisionase
Region P	SCO5605–SCO5620	46/64	72	Boundary sigma factor <i>whiG</i>
Region Q	SCO5632–SCO5644	40/44	91	Integrase, <i>korSA</i>
Region R	SCO5715–SCO5735	57/72	79	Boundary ribosomal protein, <i>bldB</i>
Region S	SCO5906–SCO5924	28/56	50	Hypothetical proteins, xylanase
Region T	SCO6372–SCO6406	82/100	82	Recombinase
Region V	SCO6607–SCO6648	62/120	52	Helicase
Region W	SCO6806–SCO6953	73/133	55	Hypothetical proteins

^a This is calculated from the available normalized gene dataset from the two microarrays

least some of the open reading frames compared to the G+C bias expected for the 1st, 2nd and 3rd codon positions of *Streptomyces* genes. Eight regions, A, B, F, I, M, O, Q and T contain transposon related genes near to or within the region. Four regions, H, N, P and R are flanked by highly conserved genes such as a ribosomal protein or sigma factor genes, which could encourage interspecific recombination. Finally, five regions consist largely of hypothetical proteins with no known similarity to any known protein as yet; these regions are G, J, L, S and W. Region L is particularly interesting as there is a central core of conserved gene flanked by two subregions that are highly not conserved. One of these genes is a putative spore septum determining protein, while the rest have unknown functions. Taken as a whole, the results suggests that *S. coelicolor* may have recently acquired all these regions either by transposition or by interspecific/intraspecific recombination (Wolf

et al. 2002; Zhang et al. 2002). It is also unlikely that they were acquired from any of the four species studied here. There are other regions that could potentially be identified as lateral transfer positions using less stringent criteria and a wider screening of genomes might help to support these additional regions as being involved in horizontal transfer. In addition, such a wider screen might allow the identification of possible origins of these regions in other species.

Gene conservation in the terminal regions of the four *Streptomyces* species

As has been mentioned earlier, the two regions at either terminus are much less well conserved than the central core region; these extend from SCO0001 to about SCO1100 on the left arm of the chromosome and from about SC7600 to SCO7845 on the right

Table 2 Genes from the (a) left terminal, (b) right terminal region of *Streptomyces coelicolor* showing microarray conservation in all four species

(a)	
SCO0002 <i>ttrA</i>	SCO0800 putative TetR-family transcriptional regulatory protein
SCO0142 hypothetical protein	SCO0802 hypothetical protein
SCO0150 hypothetical protein	SCO0810 putative ABC transporter permease
SCO0201 putative integral membrane protein	SCO0830 putative penicillin-binding protein
SCO0232 hypothetical protein	SCO0839 putative transmembrane transport protein
SCO0415 hypothetical protein	SCO0840 putative marR-family transcriptional regulator
SCO0443 hypothetical protein	SCO0854 hypothetical protein
SCO0452 putative SIR2-like regulatory protein	SCO0883 polypeptide deformylase
SCO0466 araC family transcriptional regulator	SCO0887 putative TetR-family transcriptional regulator
SCO0471 putative araC family transcriptional regulator	SCO0894 putative membrane protein
SCO0496 putative iron-siderophore permease transmembrane protein	SCO0895 RNA polymerase principal sigma factor HrdC
SCO0536 hypothetical protein	SCO0900 putative transmembrane efflux protein
SCO0538 probable sugar transporter sugar binding lipoprotein	SCO0905 putative membrane protein
SCO0544 hypothetical secreted protein	SCO0907 putative dehydrogenase
SCO0546 pyruvate carboxylase	SCO0925 putative lysR-family transcriptional regulator
SCO0551 putative histidine kinase protein	SCO0926 hypothetical protein
SCO0552 putative response regulator	SCO0931 putative secreted proline-rich protein
SCO0565 putative polyprenyl synthetase	SCO0942 putative RNA polymerase sigma factor
SCO0584 putative cytochrome	SCO0943 hypothetical protein
SCO0591 putative lysozyme precursor	SCO0947 putative integral membrane protein
SCO0592 hypothetical protein	SCO0949 hypothetical protein
SCO0614 hypothetical protein	SCO1011 conserved hypothetical protein
SCO0619 putative membrane protein	SCO1015 hypothetical protein
SCO0637 hypothetical protein	SCO1018 putative isomerase
SCO0690 possible oxidoreductase	SCO1021 hypothetical protein
SCO0695 hypothetical protein	SCO1022 hypothetical protein
SCO0701 hypothetical protein	SCO1024 hypothetical protein
SCO0707 putative branched-chain amino acid ABC transport permease	SCO1034 putative tetR-family regulatory protein
SCO0708 putative branched-chain amino acid ABC transport protein	SCO1036 putative phosphotriesterase-family protein
SCO0709 putative branched-chain amino acid transport ATP-binding protein	SCO1040 putative DNA repair protein
SCO0710 putative branched-chain amino acid transport ATP-binding protein	SCO1041 hypothetical protein
SCO0765 secreted endoglucanase	SCO1043 putative transcriptional regulatory protein
SCO0779 conserved hypothetical protein	SCO1044 putative secreted protein
SCO0788 hypothetical protein	SCO1046 putative metal transporter ATPase
SCO0790 putative hydrolase	
(b)	
SCO7649 putative two-component system sensor kinase	
SCO7677 putative secreted solute-binding protein	
SCO7678 putative metal transport integral membrane protein	
SCO7679 putative transport system integral membrane protein	

Table 2 continued

(a)

SCO7680 putative ABC transporter ATP-binding protein
SCO7681 putative AMP-binding ligase
SCO7682 putative non-ribosomal peptide synthase
SCO7684 conserved hypothetical protein
SCO7685 conserved hypothetical protein
SCO7687 putative thioesterase
SCO7688 hypothetical protein
SCO7689 putative ABC transporter ATP-binding protein
 SCO7718 hypothetical protein
 SCO7720 hypothetical protein
 SCO7724 hypothetical protein
 SCO7734 Tpg protein

Bold indicates groups of consecutive genes that may form a single transcriptional unit

arm. The boundaries of these regions are not absolutely clear-cut, but what is clear is that as one moves towards the centre of the genome, gene conservation increases beyond these points. This can be clearly seen in Fig. 3 where the gene conservation is plotted using a moving window for the four species, but it is also clear that the lack of conservation is not uniform across the terminal regions and that areas of higher gene conservation can be identified.

The significant interest in the terminal regions arises because the genomes of all *Streptomyces* that have been examined are linear and the problem of how the termini of such a molecule replicate is of particularly importance. Recent studies have indicated that two genes in particular, *tpgA* (SCO7734) and *tapA* (SCO7733), are involved in this process (Yang et al. 2002; Bao and Cohen 2001). *tpgA* encoding the terminal protein that covalently binds to the termini of many linear *Streptomyces* replicons is conserved across all four species. In *S. avermitilis* this is also true based on sequence data and, further more, there are multiple copies of *tpgA* unlike *S. coelicolor*. The signal level of the *S. avermitilis* gene at +1.2 supports the presence of these multiple copies. The signal levels for the other three species are between about −0.3 and −0.1, which supports a single slightly diverging copy of this gene in these species. However, if two copies are present then the sequence divergence may be higher. Furthermore, *tapA* is also conserved except for *S. maritimus*, which seems to be more divergent at −0.8. It should be noted that the

presence of these two genes is not a criteria for defining a genome with a linear topology, but the presence of one or both is certainly suggestive (Dary et al. 2000; Wang et al. 1999; Huang et al. 1998; Lin and Chen 1997). Finally, *ttrA* is known to be involved in chromosomal transfer and is found very close to the telomere of *S. coelicolor* and *S. avermitilis*. This is also conserved in all four species suggesting the genetic exchange is highly important in *Streptomyces* and related species.

The two terminal regions encompass the major areas that are prone to deletion in many *Streptomyces* species and are therefore not essential except for linear terminal replication and genetic exchange. Given the relatively high lack of conservation of genes in this region, genes that are present in all four species represent an interesting class. A full list of all genes conserved in all four species in the terminal regions is provided in Tables 2a and 2b. There are 36 hypothetical genes that show high similarity in the two terminal regions. Analysis of these groups of conserved genes using Artemis v7 (The Sanger Institute) identifies a total of five groups of genes that may make up possible single transcriptional units. These are SCO0551–SCO0552, SCO0705–SCO0710, SCO1021–SCO1024, SCO7677–SCO7680 and SCO7682–SCO7688. In addition to *TpgA* and *TapA*, it is possible that there are other genes involved in terminal replication and these may be among the conserved genes present in the terminal regions. Although possible candidates can

be deduced from a direct comparison of the two known *Streptomyces* genome sequences, they are many in number. Using the microarray analysis of the Actinomycetes in this study, the candidates can be reduced significantly. From candidates in Tables 2a and 2b, two possible transcriptional units seem to be potential candidates for involvement in terminal replication; these are SCO1021–SCO1024 (hypothetical proteins), and SCO7677–SCO7689 (including hypothetical proteins, an AMP-binding ligase and membrane proteins). Gene knockout studies may be able to identify possible functions for these and other gene candidates, especially the other hypothetical proteins that are conserved in these four species.

Conservation of functional groups of genes across the four *Streptomyces* species

One approach to analyzing genetic variation across these four *Streptomyces* species is to look at the functional groupings of genes. Such an approach

should allow the identification of strain versus genus specific genes especially when there are large numbers of genes with related functions such as sigma factors or where there are two copies of a gene, such as *ftsK*. However, because microarray data paints a broad picture across a whole genome, it is essential that once a gene or genes has been targeted based on microarray data, that experimental verification by other means is carried out. However, it is hoped that this dataset will be able to help researchers prioritize their gene targets better. The genes of the *S. coelicolor* chromosome have been grouped based on the scheme of M. Riley and colleagues for *E. coli* (ecocyc.org) modified for *S. coelicolor* (http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml) and we used this classification. The genes involved in ribosomal proteins synthesis and modification should be highly conserved and the results indicate that almost all of them are present in all four species (Table 3; Supplementary Fig. 4). The only exceptions are SCO0436, SCO0509 SCO3430 and

Table 3 Microarray data for ribosomal proteins from the four species

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO0436 probable 50S ribosomal protein	−0.35	−0.13	0.44	−0.29
SCO0569 putative 50S ribosomal protein fragment	−0.75	0.63	−0.42	0.27
SCO1150 50S ribosomal protein L31	0.56	−0.47	−0.24	0.14
SCO1505 30S ribosomal protein S4	0.36	−0.35	0.76	−0.31
SCO1598 50S ribosomal protein L20	1.50	0.63	0.89	0.34
SCO1599 50S ribosomal protein L35	0.23	0.41	0.49	−0.51
SCO1998 30S ribosomal protein S1	1.39	0.77	0.99	0.91
SCO2563 30S ribosomal protein S20	−0.34	−0.34	0.77	0.33
SCO2596 50S ribosomal protein L27	1.01	0.59	0.08	0.82
SCO2597 ribosomal protein L21	0.27	0.08	0.64	−0.18
SCO3124 ribosomal L25p family protein	0.39	0.31	−0.23	−0.89
SCO3427 putative 50S ribosomal protein L31	0.24	0.37	0.22	0.60
SCO3428 putative 50S ribosomal protein L33	0.15	0.28	0.54	0.09
SCO3429 putative 50S ribosomal protein L28	0.68	0.16	0.55	0.45
SCO3430 putative 30S ribosomal protein S14	−0.80	0.10	−0.17	−0.19
SCO3880 putative 50S ribosomal protein L34	1.02	0.13	0.24	0.71
SCO3906 putative 30S ribosomal protein S6	0.70	0.94	1.11	−0.18
SCO3909 putative 50S ribosomal protein L9	1.30	0.01	0.87	−1.27
SCO4648 50S ribosomal protein L11	1.67	0.43	0.90	0.47
SCO4649 50S ribosomal protein L1	0.62	0.53	−0.25	0.92
SCO4652 50S ribosomal protein L10	0.42	−0.43	0.64	−0.35
SCO4653 50S ribosomal protein L7/L12	1.22	1.03	0.79	0.45
SCO4659 30S ribosomal protein S12	0.74	0.65	0.70	−0.53

Table 3 continued

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO4660 30S ribosomal protein S7	0.57	−0.23	0.68	0.12
SCO4701 30S ribosomal protein S10	1.19	1.17	1.16	−0.21
SCO4702 50S ribosomal protein L3	0.92	0.02	0.84	0.49
SCO4703 50S ribosomal protein L4	1.16	0.91	0.59	0.23
SCO4704 50S ribosomal protein L23	0.85	1.44	1.24	0.36
SCO4705 50S ribosomal protein L2	0.85	−0.12	0.84	0.22
SCO4706 30S ribosomal protein S19	0.06	0.24	0.32	−0.26
SCO4707 50S ribosomal protein L22	0.96	0.69	0.64	0.15
SCO4708 30S ribosomal protein S3	1.15	0.38	0.78	1.07
SCO4709 50S ribosomal protein L16	0.52	0.67	1.09	1.26
SCO4710 50S ribosomal protein L29	0.33	−0.06	−0.19	0.41
SCO4711 30S ribosomal protein S17	0.59	0.92	0.51	−0.13
SCO4712 50S ribosomal protein L14	1.05	0.35	0.48	0.82
SCO4713 50S ribosomal protein L24	1.09	0.88	0.63	0.64
SCO4714 50S ribosomal protein L5	1.24	0.78	0.77	1.03
SCO4715 30S ribosomal protein S14	0.39	0.03	0.19	0.12
SCO4716 30S ribosomal protein S8	1.18	−0.11	0.64	−0.76
SCO4717 50S ribosomal protein L6	0.96	0.82	0.79	−0.02
SCO4718 50S ribosomal protein L18	0.09	0.30	0.57	0.74
SCO4719 30S ribosomal protein S5	1.56	0.66	1.02	−0.09
SCO4720 50S ribosomal protein L30	0.13	0.39	0.64	0.26
SCO4721 50S ribosomal protein L15	1.79	0.42	0.80	0.84
SCO4726 50S ribosomal protein L36	0.46	−0.13	0.35	−0.10
SCO4727 30S ribosomal protein S13	0.63	−0.23	0.62	−0.17
SCO4728 30S ribosomal protein S11	1.12	0.55	0.87	0.27
SCO4730 50S ribosomal protein L17	0.69	0.27	0.86	0.51
SCO4734 50S ribosomal protein L13	−0.27	0.45	0.50	0.40
SCO4735 30S ribosomal protein S9	0.36	−0.02	0.00	−0.01
SCO5359 50S ribosomal protein L31	0.86	0.22	1.40	0.46
SCO5564 putative 50S ribosomal protein L28	0.60	0.28	0.21	0.51
SCO5591 30S ribosomal protein S16	0.44	0.03	0.57	−0.60
SCO5595 50S ribosomal protein L19	0.77	0.78	1.54	−0.03
SCO5624 30S ribosomal protein S2	1.16	0.48	1.52	0.30
SCO5736 30S ribosomal protein S15	0.70	0.41	0.79	−0.46
Mean hybridization score for ribosomal protein genes	0.67	0.35	0.61	0.18

Bold values indicate that the signal for that gene is more than 2SD below the mean core signal for that species and such a value is suggestive of either gene absence or very low similarity

SCO3909 in *S. avermitilis* and SCO4716 and SCO5514 in *K. aureofaciens*. Of these genes, SCO0436, SCO0509 and SCO5514 represent duplicate genes in the *S. coelicolor* genome and therefore the choice of the microarray sequence will have had a significant effect on the heterologous hybridization. There is no obvious explanation for the failure to

hybridize of the other two genes, but as a whole, this dataset supports the integrity of the array system for analysis of genome content as these genes are scattered across the whole *Streptomyces* genome.

Table 4 shows genes identified as possible sigma factors, anti-sigma factors and anti-sigma factor antagonists. The genes found in the central core

Table 4 Conservation across the four species of genes annotated as sigma factors or related proteins in *Streptomyces coelicolor*

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>	
SCO0037 putative sigma factor	−1.04	−0.90	−0.88	−1.49	
SCO0159 putative ECF sigma factor	−1.26	−0.65	2.05	−0.60	
SCO0194 putative sigma factor	−0.87	−0.35	−0.61	−0.52	
SCO0255 putative transcriptional regulator	−0.64	−0.37	−0.99	−0.46	
SCO0414 putative RNA polymerase sigma factor	−0.05	−0.28	−0.15	−0.22	Conserved
SCO0598 putative anti anti sigma factor	0.11	0.50	−0.08	0.57	Conserved
SCO0599 putative regulator of sig8	−1.41	−1.07	−0.84	−1.08	
SCO0632 putative RNA polymerase sigma factor	−0.14	0.19	−0.82	0.11	
SCO0672 putative anti-sigma factor antagonist	−0.10	−0.40	−0.19	0.12	
SCO0781 putative anti sigma factor antagonist	−0.79	−0.86	−1.06	−0.83	
SCO0803 putative RNA polymerase sigma factor	−0.25	−0.09	−0.51	−0.01	
SCO0864 probable ECF-family sigma factor	−0.74	−0.86	−1.04	−0.50	
SCO0866 probable ECF-family sigma factor	−0.13	0.19	−0.28	0.23	Conserved
SCO0869 putative anti-sigma factor antagonist	−0.59	−0.90	−1.23	−0.90	
SCO0895 RNA polymerase principal sigma factor HrdC	0.52	0.43	1.12	0.34	Conserved
SCO0942 putative RNA polymerase sigma factor	0.34	0.81	0.61	0.45	Conserved
SCO1263 putative ECF-sigma factor	−0.17	−0.26	0.08	0.35	Conserved
SCO1276 RNA polymerase ECF sigma factor	−1.22	−0.55	0.60	−0.88	
SCO1564 putative RNA polymerase sigma factor	1.04	−0.33	1.34	−0.47	
SCO1723 putative RNA polymerase sigma factor	0.19	−0.39	−0.15	0.52	Conserved
SCO1876 putative RNA polymerase sigma factor	−1.01	−0.93	−0.77	−0.50	
SCO2465 RNA polymerase principal sigma factor	0.76	0.74	0.97	0.64	Conserved
SCO2639 putative RNA polymerase sigma factor	0.74	0.22	0.01	0.27	Conserved
SCO2954 putative RNA polymerase sigma factor	1.12	−0.51	0.94	−0.46	
SCO3066 putative regulator of Sig15	0.53	0.29	0.88	0.31	Conserved
SCO3067 putative anti anti sigma factor	−0.74	−1.25	0.89	−0.42	
SCO3068 putative RNA polymerase sigma factor	0.33	−0.06	1.07	−0.41	Conserved
SCO3202 RNA polymerase principal sigma factor	0.98	0.29	1.40	0.19	Conserved
SCO3323 putative RNA polymerase sigma factor	0.76	0.49	1.08	0.27	Conserved
SCO3356 ECF sigma factor 37	−0.05	0.11	0.73	0.46	Conserved
SCO3450 putative RNA polymerase sigma factor (ECF subfamily)	0.16	−0.09	−0.79	−0.23	
SCO3548 putative anti-sigma factor	−0.57	0.50	0.49	−0.06	
SCO3549 bldG putative anti-sigma factor antagonist	−0.03	−0.16	0.21	−0.20	Conserved
SCO3613 putative RNA polymerase sigma factor	0.57	0.15	0.08	0.46	Conserved
SCO3692 putative anti-sigma factor antagonist	0.14	0.64	−0.27	0.13	Conserved
SCO3709 putative ECF sigma factor	0.02	0.06	−0.21	0.61	Conserved
SCO3715 putative ECF sigma factor	0.45	0.92	−0.28	−0.27	Conserved
SCO3736 putative RNA polymerase ECF sigma factor	−0.06	−0.29	0.28	−0.09	Conserved
SCO3892 putative RNA polymerase sigma factor	0.68	0.21	0.80	−0.27	Conserved
SCO4027 putative anti sigma factor antagonist	−0.04	1.11	−0.59	0.58	
SCO4034 putative RNA polymerase sigma factor	0.98	1.27	1.26	0.22	Conserved
SCO4035 RNA polymerase sigma factor (fragment)	1.04	1.27	0.88	0.45	Conserved
SCO4146 putative ECF subfamily sigma factor	−0.23	0.53	0.58	0.16	Conserved

Table 4 continued

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>	
SCO4409 putative RNA polymerase sigma factor	−0.10	0.12	0.10	0.64	Conserved
SCO4410 putative anti anti sigma factor	−0.89	0.07	0.16	−0.81	
SCO4452 putative sigma factor	−0.17	−0.21	−0.08	0.27	Conserved
SCO4769 ECF sigma factor	0.09	−0.38	0.68	0.61	
SCO4864 putative ECF sigma factor	0.02	−0.34	−0.12	−0.73	
SCO4866 putative ECF sigma factor	0.12	0.19	0.09	0.38	Conserved
SCO4895 putative ECF sigma factor	−0.32	−1.15	−0.11	−0.56	
SCO4938 putative ECF-sigma factor	0.17	0.43	0.24	0.64	Conserved
SCO4960 possible sigma factor	−0.04	0.05	−0.66	0.60	
SCO4996 putative RNA polymerase ECF sigma factor	−0.54	−0.04	0.58	0.48	
SCO5147 putative ECF-subfamily sigma factor	−0.39	0.35	0.79	0.59	
SCO5217 anti-sigma factor	−0.47	−0.05	0.28	−0.80	
SCO5244 anti-sigma factor	−0.32	−0.54	−0.37	−0.29	
SCO5386 putative anti-sigma factor antagonist	0.15	0.37	0.00	−0.07	Conserved
SCO5621 RNA polymerase sigma factor WhiG	0.79	0.92	0.64	−0.27	Conserved
SCO5820 hrdB, major vegetative sigma factor	1.36	1.06	1.53	1.09	Conserved
SCO5934 putative sigma factor	0.07	0.25	−0.58	0.17	
SCO6239 putative sigma factor	−0.92	−1.27	−1.84	−0.74	
SCO6996 putative RNA polymerase sigma factor	−0.34	0.29	0.00	−0.04	Conserved
SCO7099 putative RNA polymerase sigma factor	−0.20	−0.38	0.38	0.29	
SCO7104 putative RNA polymerase sigma factor	−0.70	−0.02	−0.79	0.56	
SCO7112 putative ECF-family RNA polymerase sigma factor	−0.35	−0.25	−1.65	−0.38	
SCO7144 putative ECF sigma factor	−0.62	0.13	−0.91	0.34	
SCO7314 probable RNA polymerase sigma factor	−0.25	−0.11	0.23	0.46	Conserved
SCO7323 anti-sigma factor antagonist	0.30	−0.01	0.23	0.31	Conserved
SCO7325 anti-sigma factor antagonist	−0.37	−0.68	−0.23	−1.15	
SCO7341 putative RNA polymerase secondary sigma factor	−0.09	0.54	0.29	0.44	Conserved
SCO7573 putative anti-sigma factor antagonist	−0.18	0.03	−1.69	0.21	
SCO7619 putative anti sigma factor antagonist	−0.28	0.37	−0.97	−0.64	
SCO7754 putative anti-sigma factor antagonist	−1.20	−0.51	−1.89	0.02	
Mean hybridization score for ribosomal protein genes	0.01	0.06	0.10	0.05	NA

Bold values indicate that the signal for that gene is more than 2SD below the mean core signal for that species and such a value is suggestive of either gene absence or very low similarity. A conserved gene is one that seems to be present in all four species. NA, Not applicable

region are more conserved. As would be expected, the major sigma factors such as hrdA, hrdB, hrdC and hrdD are conserved as well as many of the other studied sigma factors of *S. coelicolor* such as are sigA, sigE, sigF, sigG, sigR, sigT and whiG. Overall, fewer regulation genes from this group (anti-sigma factors and anti-anti-sigma factors) are conserved than sigma factors themselves. This analysis allows the identification of new candidate sigma factors for

further study outside of the well studied ones, but within *S. coelicolor* and in other species. Overall, the results support the hypothesis that there is a core of sigma factors essential to keeping protein synthesis in *Streptomyces* running smoothly. The functionality of the rest may vary and include complete silence of some gene fragments, duplication of function, involvement in specific secondary metabolic activities and species/genus specific functions.

All four species studied here undergo differentiation and spore formation and as such would be expected to retain most genes involved in cell division/sporulation/differentiation. This is supported by Table 5. *K. aureofaciens* shows greater gene

divergence for certain genes when compared to the three *Streptomyces* species and these are specifically ftsI (SCO2090) and a putative cell division protein (SCO2968). However, in general, the same genes in all four species show a higher divergence, for

Table 5 Conservation across the four species of genes in *Streptomyces coelicolor* annotated as involved in cell division, sporulation and differentiation

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO0409 sapA spore-associated protein precursor	−1.99	−1.39	−0.59	−0.67
SCO1454 putative amino oxidase	1.00	0.45	1.18	−0.17
SCO1489 bldD putative DNA binding protein	0.99	0.76	1.02	0.58
SCO1772 putative partitioning or sporulation protein	0.69	0.35	0.39	0.54
SCO2082 ftsZ cell division protein	1.44	0.89	0.97	0.95
SCO2083 ftsQ sporulation protein	0.32	0.74	0.06	0.26
SCO2084 murG	0.86	−0.01	0.49	0.32
SCO2085 fts W putative cell division protein	0.82	0.81	0.59	0.50
SCO2086 murD	0.58	0.23	0.35	0.45
SCO2087 murX	0.41	0.05	0.51	−0.30
SCO2088 murF	1.18	0.73	0.67	0.01
SCO2089 murE	0.73	0.47	0.36	0.31
SCO2090 ftsI cell division protein	0.80	0.01	0.45	−0.50
SCO2607 Sfr protein	0.73	0.73	0.91	−0.01
SCO2608 penicillin binding protein	−0.04	−0.25	0.45	−0.92
SCO2609 mreD rod shape-determining protein	0.09	0.07	0.66	0.43
SCO2610 mreC rod shape-determining protein	−0.42	−0.16	0.17	−0.52
SCO2611 mreB rod shape-determining protein	0.98	0.80	0.68	0.19
SCO2620 putative cell division trigger factor	0.81	−0.12	0.50	0.25
SCO2968 putative cell division protein	0.35	−0.17	0.34	−0.39
SCO2969 ftsE cell division ATP-binding protein	0.37	−0.43	0.00	−0.51
SCO3034 whiB sporulation regulatory protein	0.23	0.50	0.81	0.26
SCO3323 bldB putative RNA polymerase sigma factor	0.76	0.49	1.08	0.27
SCO3404 ftsH2 cell division protein ftsH homolog	1.11	0.51	1.07	0.15
SCO3549 bldG putative anti-sigma factor antagonist	−0.03	−0.16	0.21	−0.20
SCO3557 putative septum site determining protein	0.31	0.45	−0.19	0.92
SCO3558 putative morphological differentiation-associated protein	0.69	−0.26	1.62	−0.20
SCO3846 putative FtsW/RodA/SpoVE family cell cycle protein	1.11	0.31	1.07	0.61
SCO3886 putative partitioning or sporulation protein	0.00	−0.83	−0.43	−0.98
SCO3887 putative partitioning or sporulation protein	−0.19	−0.19	0.24	−1.04
SCO3934 ftsK/spoIIIE family protein	−0.56	0.39	−1.18	−0.53
SCO4014 sporulation associated protein	−0.87	−0.93	−0.81	−1.17
SCO4184 mfc aerial mycelium formation	0.12	0.17	0.01	−0.05
SCO4508 putative cell division-related protein	−0.62	−0.27	0.07	0.59
SCO4531 putative septum determining protein	−0.56	−0.88	0.12	−0.09
SCO4620 traB1 putative sporulation-related protein	−0.49	−0.68	0.19	−0.34
SCO4621 traA1 putative sporulation-related protein	−0.03	−0.51	1.33	−0.38
SCO4767 putative regulatory protein	0.14	0.00	1.67	−0.01

Table 5 continued

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO4768 bldM putative two-component regulator	1.06	1.00	0.92	0.61
SCO5006 minD1 putative septum site-determining protein	−0.31	−0.04	0.58	−0.28
SCO5008 minD3 putative septum site-determining protein	0.04	−0.21	−0.11	−0.07
SCO5112 BldKA	−0.42	0.92	−0.75	0.51
SCO5114 BldKC	−0.39	−0.08	−1.23	−0.23
SCO5115 BldKD	0.01	−0.18	−0.89	−0.15
SCO5116 bldKE putative peptide transport system ATP-binding protein	−0.04	−0.22	−0.70	−0.03
SCO5314 whiE protein VII	−1.24	−0.23	0.03	−0.67
SCO5315 polyketide cyclase	−0.39	−0.82	−0.31	−0.25
SCO5316 acyl carrier protein	−0.42	−0.29	0.75	0.03
SCO5318 polyketide beta-ketoacyl synthase alpha	−0.03	−0.13	0.86	0.43
SCO5321 polyketide hydroxylase	0.12	−0.09	1.50	0.27
SCO5587 ftsH cell division protein FtsH homolog	−0.05	0.31	0.23	0.21
SCO5621 whiG RNA polymerase sigma factor WhiG	0.79	0.92	0.64	−0.27
SCO5723 bldB putative regulator, BldB	−1.47	−0.76	−1.27	−1.05
SCO5750 ftsK homolog	0.67	0.16	0.65	2.52
SCO5819 whiH, sporulation transcription factor	0.68	0.12	0.16	0.13
SCO6029 whiI two-component regulator	0.14	−0.26	0.77	0.92
Mean hybridization score for ribosomal protein genes	0.05	−0.05	0.27	−0.02

Bold values indicate that the signal for that gene is more than 2SD below the mean core signal for that species and such a value is suggestive of either gene absence or very low similarity

example *sapA*, which is a protein associated with the spore surface hydrophobicity. As spore morphology varies a lot in the *Streptomyces*, high variability/gene loss in such a gene is not unexpected. Other genes that show higher divergence are those involved in partitioning and cell division. This suggests that the genes and thus the proteins involved in these functions may differ from species to species in order to create the variation seen in aerial mycelium and spore structure across *Streptomyces* species. Specifically, SCO3934, an *ftsK* family protein gene is less well conserved than its homologue. This suggests that SCO5750 may produce the major ftsK protein. Other Fts proteins show a similar pattern with at least one homologue being well conserved. This may well help an understanding of the relationships between the genes involved in cell division and will allow better identification of specific targets for further study. One anomaly that stands out is *bldB*. This gene consistently shows a low level of hybridization. A comparison of the *bldB* gene sequence between *S. coelicolor* and *S. avermitilis* shows a nucleotide identity of about 87%, which ought to give a signal in

the region of 0.0 or better. As two different arrays are used in this study, mechanical problems with this spot can probably be eliminated as the source of the anomaly. We suggest that because this is a relatively small gene, the PCR product chosen for both arrays may be the reason for this result. This emphasizes that array data should be used with a degree of caution and needs to be backed up by other experimental evidence when specific genes are being investigated.

The genes involved in DNA replication, repair, restriction/modification are shown in Table 6 and only about 20% of these genes are not conserved relatively well across all four species. This is to be expected as DNA replication and repair are core functions. Most of the genes that show higher levels of gene divergence are found in the terminal regions of the linear chromosome and probably are genes that perform functions that are not essential to cell survival because the terminal regions of *Streptomyces* chromosomes are unstable and liable to deletion without lethality. Of particular interest are SCO0183 and SCO0842 (deoxiribopyrimidine photolyases);

Table 6 Conservation across the four species of genes in *Streptomyces coelicolor* annotated as involved in DNA replication, repair, restriction and modification

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO0183 putative deoxyribodipyrimidine photolyase	−1.38331	−0.76404	−1.15786	−0.78973
SCO0760 putative methyltransferase	−0.23598	−0.17603	0.079613	−0.16712
SCO0842 putative deoxyribodipyrimidine photolyase	0.002475	0.167429	−0.05198	0.35453
SCO0918 putative excinuclease ABC subunit A	−0.28707	−0.34813	0.010852	−0.46861
SCO0945 putative formamidopyrimidine-DNA glycosylase	−0.34857	−0.60698	−0.4371	−0.38083
SCO1040 putative DNA repair protein	0.047315	0.568723	0.456168	0.367071
SCO1050 putative DNA protection protein	−0.47479	0.093299	0.451487	−0.92697
SCO1114 uracil DNA glycosylase	−0.34573	0.357182	0.433902	0.950894
SCO1167 putative helicase (fragment)	0.604813	0.072648	−0.40455	−0.25074
SCO1180 putative DNA polymerase III beta chain	−0.347	−0.21511	−0.42838	−0.20922
SCO1202 putative DNA ligase	0.308497	0.202188	−0.15524	0.096586
SCO1203 putative MutT-like protein	−0.17791	−0.3705	0.324233	−0.28871
SCO1255 G/U mismatch-specific DNA glycosylase	0.521679	0.41904	0.321208	0.429148
SCO1343 uracil-DNA glycosylase	0.61294	0.00339	0.311478	0.036821
SCO1380 putative DNA damage inducible protein	0.763064	0.313948	0.214931	0.677262
SCO1395 mutT-like protein	0.068215	0.572527	0.17071	0.414092
SCO1475 putative primosomal protein	0.048892	0.689232	0.53545	0.457655
SCO1518 ruvB holliday junction DNA helicase	1.136489	0.638803	0.930067	1.045593
SCO1519 ruvA holliday junction DNA helicase	−0.57721	−0.23275	−0.15374	0.008274
SCO1520 ruvC crossover junction endodeoxyribonuclease	1.079786	0.708131	0.886363	0.979821
SCO1534 putative DNA polymerase III	0.3296	0.371242	−0.10025	0.090739
SCO1739 putative DNA polymerase III	1.128049	0.423158	1.111964	0.326701
SCO1780 putative DNA repair protein	−0.14578	−0.09175	0.258713	−0.32746
SCO1792 putative 3-methyladenine DNA glycosylase	−0.20485	−0.07755	0.083499	−0.2824
SCO1827 putative DNA polymerase III	0.559715	0.711265	0.512873	0.343849
SCO1966 ABC excision nuclease subunit B	0.047382	0.06896	1.085903	−0.05934
SCO1969 putative DNA-methyltransferase	−0.19025	−0.12184	0.143367	0.56545
SCO2003 DNA polymerase I	1.172176	0.188493	0.498126	0.201411
SCO2468 DNA primase	0.82715	0.520552	1.25613	0.69797
SCO2626 putative DNA repair hydrolase (fragment)	0.289916	0.337365	0.567783	0.326821
SCO2863 putative helicase	−0.47935	−0.68124	−1.92881	−0.81052
SCO2952 putative helicase protein	0.513952	0.38864	0.750517	0.379556
SCO3109 putative transcriptional-repair coupling factor	−0.87255	−0.29611	0.759684	−0.33637
SCO3351 putative DNA repair protein	−0.95043	−0.68521	0.24637	−0.48466
SCO3352 putative DNA-binding protein	0.090569	−0.09602	0.536392	−0.2275
SCO3434 putative DNA polymerase I	0.8554	1.651768	0.874246	1.328528
SCO3510 putative DNA methylase	0.402433	−1.11393	−2.15274	−2.07607
SCO3541 putative DNA polymerase	0.003597	−0.40289	0.297076	−0.75003
SCO3543 probable DNA topoisomerase I	0.798705	1.33477	0.87884	1.226661
SCO3550 putative helicase	0.263644	0.13935	0.559872	0.216127
SCO3618 putative recombination protein	0.533404	0.052623	0.511469	−0.36893
SCO3873 DNA gyrase subunit A	1.293458	−0.20851	0.985546	−1.2494
SCO3874 DNA gyrase subunit B	0.99333	0.958247	1.250669	0.37466
SCO3878 DNA polymerase III	0.05375	0.026053	0.328502	−0.62343

Table 6 continued

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO3879 chromosomal replication initiator protein (fragment)	1.25983	1.572531	0.731065	−0.12482
SCO4092 ATP-dependent helicase	−0.00752	0.055194	0.978098	−0.21188
SCO4143 putative mutT-like protein	0.021035	0.435471	0.479161	−0.28423
SCO4272 putative mutT-like protein	−0.1352	0.016206	−0.32968	0.527274
SCO4351 putative DNA invertase	−1.26688	−0.78144	−1.29284	−0.68489
SCO4495 putative DNA polymerase related protein	−0.09624	−0.75937	0.582419	0.316259
SCO4577 putative helicase	0.726962	0.053196	1.044155	−0.09189
SCO4797 putative ATP-dependent DNA helicase II	0.323366	0.286421	0.749868	0.026912
SCO5064 putative bifunctional protein	−0.20262	−0.49259	−1.90763	−0.29433
SCO5143 DNA-3-methyladenine glycosylase I	−0.86594	0.342489	0.276611	0.981316
SCO5183 putative ATP-dependent DNA helicase	0.235633	0.120765	0.767504	0.591935
SCO5184 putative ATP-dependent DNA helicase	0.180638	−0.17587	0.208839	0.666073
SCO5188 putative ATP-dependent DNA helicase	0.204048	−0.36005	0.808168	−0.04087
SCO5331 putative DNA methylase	−1.52864	−2.37483	−3.76436	−2.18209
SCO5494 putative DNA ligase	0.124781	0.288429	0.215946	−0.0149
SCO5566 putative ATP-dependent DNA helicase	0.408995	1.186714	0.428183	0.825946
SCO5567 putative methylase	0.338198	−0.82207	0.596196	−0.83241
SCO5573 formamidopyrimidine-DNA glycosylase	0.387552	−0.08353	0.601585	0.435209
SCO5760 DNA glycosylase	0.846985	0.02235	0.675876	0.46121
SCO5770 RecX protein	−0.1193	−0.72104	−0.58075	0.079675
SCO5802 putative ATP-dependent helicase	0.823429	−0.10433	0.511595	0.058287
SCO5803 SOS regulatory protein LexA	0.143322	−0.78751	−0.0905	−0.16946
SCO5805 ribonucleotide reductase	0.235182	0.271514	0.98906	−0.06823
SCO5815 probable ATP-dependent DNA helicase	−0.52023	−0.31138	−0.75254	−0.71096
SCO5822 gyrB2, probable DNA gyrase	0.167275	0.223473	0.426849	0.393883
SCO5836 DNA gyrase-like protein	0.725708	0.507498	0.937531	0.073608
SCO6084 putative DNA polymerase	−0.05381	0.165634	0.084564	−0.16407
SCO6151 putative methylated-DNA-protein-cysteine methyltransferase	−0.88705	−0.08112	0.375214	0.207725
SCO6262 putative helicase 6884138:6887071 forward MW:103912	0.260624	−0.28836	0.753508	−0.08897
SCO6405 putative DNA recombinase	−0.16958	−0.65562	−1.15936	−0.30768
SCO6462 putative methylated-DNA-protein-cysteine methyltransferase	−0.09961	−0.00285	0.231836	0.026866
SCO6640 putative ATP-dependent helicase	−0.56659	−0.52367	−1.18266	−0.38994
SCO6707 putative DNA ligase	−0.25409	0.618121	−0.0252	−0.3333
SCO6844 putative DNA methylase.	0.487806	0.475711	−0.59387	0.230037
SCO6907 putative DNA ligase.	−0.71491	0.181384	−0.67564	−0.69734
SCO7345 probable ATP-dependent DNA ligase	0.176692	0.207606	0.340255	0.085159
SCO7522 putative DNA ligase	−0.31874	−0.0533	−0.63701	0.470458
Mean hybridization score	0.113633	0.007441	0.153752	−0.04491

Bold values indicate that the signal for that gene is more than 2SD below the mean core signal for that species and such a value is suggestive of either gene absence or very low similarity

these repair system would seem to be absent in *S. lividans* and *S. maritimus*, but a homologue is present in *S. avermitilis* (confirmed by the genome sequence) and in *S. cattleya*. This confirms the high

variability found for this repair function across the *Streptomyces* (Kobayashi et al. 1989). A similar situation of high variability is found for the *mutT* homologues, potential 8 hydroxy-dGTP hydrolases.

Knockout of this gene has been shown to increase the A:T to G:T mutation rate and thus it has a possible repair function (Kamiya et al. 2004). The genes for *recA* (SCO5769), *recF* (SCO3876) and *recR* (SCO3618) are present in all four species; however, the *recX* (SCO5770), is more divergent and gives a low signal for *S. cattleya* and *S. maritimus*. SCO6405, a putative DNA recombinase, is scored as absent in all four species suggesting that there is redundancy in the *Streptomyces* genes concerned with recombination or that this gene is transposon related. The latter is supported by low homology to *S. avermitilis* putative integrases/recombinases. There are four genes encoding DNA gyrases on the microarray, namely, *gyrA* DNA gyrase subunit A (SCO3873) and *gyrB* DNA gyrase subunit B (SCO3874) together with SCO5836 and SCO5822 and these may be TopIV homologues involved in resolving chromosome concatenates. All are conserved although the conservation of SCO5822 *gyrB2* is lower. Thus both sets of gyrase genes would seem to be important. As expected, SCO1518, a *ruvB* Holliday junction protein gene and SCO1520, a *ruvC* crossover junction endonuclease are conserved across all the species. Unexpectedly, although probably present in all species, SCO1519 *ruvA* is much more divergent than the other two genes in this Holliday junction complex.

This diversity is unexpected and not easily explicable except by the fact that recombination in *Streptomyces* may occur via a more variable mechanism than in other groups of bacteria and this is then reflected in the greater divergence of SCO1519 *ruvA*. All three genes annotated as a DNA polymerase I homologue are conserved as are four out of the five DNA polymerase III homologues, suggesting that there are roles for all of these conserved genes in *Streptomyces*. Two other unclassified DNA polymerase type genes, SCO4495 and SCO6084 are also conserved and thus may have important functions. There is, however, more diversity among the helicases and methylases/methyltransferases. With the helicases, three out of 14 show significant divergence and therefore most of the helicases probably have important cellular roles. Four out of nine methylases/methyltransferases show divergence. As some of these genes may be involved in the DNA modification part of restriction/modification, such diversity across strains is not unexpected. Finally, four out of six ligases show divergence, perhaps reflecting the fact that the origin of a number of these ligases might be from bacteriophages.

Table 7 shows the genes involved in peptidoglycan and teichoic acid synthesis. In this area of metabolism, there is also a relatively high level of

Table 7 Conservation across the four species of genes in *Streptomyces coelicolor* annotated as involved in peptidoglycan biosynthesis

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO0237 putative oxidoreductase	−0.07526	−1.14854	−0.09821	−0.57347
SCO0286 putative peptidoglycan binding protein	−0.9459	−0.93949	−1.83967	−1.02755
SCO0830 putative penicillin-binding protein	0.243458	0.081341	0.556096	0.08585
SCO0936 putative oligosaccharide deacetylase	−0.78759	−1.38608	−0.58892	−1.00821
SCO1018 putative isomerase	0.390534	0.519204	0.392605	0.600019
SCO1875 putative secreted penicillin binding protein	−0.44831	0.039598	0.320356	0.256976
SCO2084 murG	0.85602	−0.00677	0.485989	0.319639
SCO2085 putative cell division protein	0.816624	0.80742	0.592214	0.502565
SCO2086 murD	0.580506	0.225348	0.347674	0.449973
SCO2087 murX	0.405731	0.049645	0.509959	−0.3047
SCO2088 murF	1.175078	0.730925	0.667383	0.006107
SCO2089 murE	0.734068	0.470869	0.363458	0.308295
SCO2345 putative peptidoglycan-binding membrane protein	−0.0329	0.016338	0.133355	0.069404
SCO2451 putative rod shape-determining protein	0.604327	0.180028	0.612243	0.205981
SCO2589 putative glycosyl transferase	0.093285	−0.41217	0.478809	−0.09528
SCO2590 putative glycosyltransferase	0.069933	−0.52938	−1.18523	−0.89326

Table 7 continued

	<i>S. avermitilis</i>	<i>S. cattleya</i>	<i>S. maritimus</i>	<i>K. aureofaciens</i>
SCO2608 penicillin binding protein	−0.0413	−0.24681	0.445143	− 0.92429
SCO2609 rod shape-determining protein	0.085879	0.06721	0.663609	0.427758
SCO2610 rod shape-determining protein	− 0.41695	−0.15798	0.168038	− 0.52083
SCO2611 rod shape-determining protein	0.979884	0.796403	0.675204	0.186331
SCO2706 putative transferase	0.233423	0.505652	− 0.55544	0.407847
SCO2707 putative transferase	0.014126	− 0.37225	0.088105	−0.12772
SCO2897 probable penicillin-binding protein	0.576646	0.409388	0.693237	0.120551
SCO2949 murA	0.418363	0.209024	0.335159	0.062475
SCO3580 putative transpeptidase	−0.09884	− 0.76666	− 1.13347	0.212152
SCO3595 putative D-alanine:D-lactate ligase	− 0.77707	− 1.21402	− 1.61266	− 1.31715
SCO3596 putative D-alanine:D-alanine dipeptidase	− 1.47294	− 0.65926	− 1.4814	− 0.85496
SCO3811 putative D-alanyl-D-alanine carboxypeptidase	− 0.50714	− 0.58617	0.155077	− 0.78413
SCO3847 putative penicillin-binding protein	0.128622	− 0.43643	0.640849	−0.18456
SCO3901 putative penicillin-binding protein	− 0.59684	− 0.78306	−0.26692	−0.39905
SCO4013 putative penicillin binding protein	−0.09095	−0.01844	−0.04797	0.164779
SCO4132 putative secreted transglycosylase	0.237226	−0.17562	−0.03522	−0.2261
SCO4439 putative D-alanyl-D-alanine carboxypeptidase	− 0.74865	− 0.93049	0.177585	− 1.21329
SCO4643 murB	−0.01505	−0.20659	0.197283	−0.35551
SCO5039 putative penicillin-binding protein	0.590106	−0.70926	0.891819	0.313533
SCO5301 putative penicillin-binding protein	−0.19347	−0.13131	0.547233	−0.22028
SCO5365 putative transferase	1.11236	−0.27334	0.569152	− 1.16844
SCO5467 muramoyl-pentapeptide carboxypeptidase	−0.15448	0.387477	−0.49044	−0.02243
SCO5560 D-alanine-D-alanine ligase	0.728112	0.209602	0.167769	0.330255
SCO5998 murA2	0.800726	0.792975	1.291368	0.051102
SCO6060 putative UDP- <i>N</i> -acetylmuramoyl-L-alanine ligase	0.19941	0.377004	0.817329	0.527531
SCO7050 putative D-alanyl-D-alanine carboxypeptidase	0.605949	0.70203	0.343827	0.524521
Mean hybridization score	0.125637	−0.10744	0.118866	−0.14494

Bold values indicate that the signal for that gene is more than 2SD below the mean core signal for that species and such a value is suggestive of either gene absence or very low similarity

conservation of genes, particularly the *murA*, *murA2*, *murB*, *murD*, *murE*, *murF*, *murG* and *murX* genes. Also conserved are the shape-determining genes SCO2609, SCO2610 and SCO2611, which may form an operon. This probably represent a core of genes together with the genes involved in biosynthesis of the cell wall that are needed to give a basic structure to the cells of any *Streptomyces* species. The penicillin binding proteins show a higher degree of variability, except for SCO2897, SCO4013 and SCO5301. The peptidases SCO3580, SCO3596, SCO3011 and SCO4439 and the D-alanine:D-lactate ligase SCO3595 all show a low level of gene conservation, perhaps because they are involved in

relatively broad cellular functions and not under a great deal of selective pressure.

Conserved genes with no known function

Genes with no known function and no homologue outside of *S. avermitilis* that are conserved across the other three *Streptomyces* species should represent genes important to specifically being a myceliate Actinobacteria and the phenotype of gene knockout strains for these genes will be particularly interesting in terms of *Streptomyces* biology. Based on the dataset here, 936 genes can be identified as annotated as either conserved hypothetical genes or non-conserved

hypothetical genes and these are shown in Supplementary Table 1. The proportion of these genes that are conserved across all four species are 9%, 20%, 13%, 16% and 12% for the left terminal region, left *Streptomyces* specific region, core region, right *Streptomyces* specific region and right terminal region, respectively. There is also a low frequency of conserved hypothetical genes in the left terminal region and right *Streptomyces* specific region, 0.78% and 0.96%, respectively compared to 3.4% for the left *Streptomyces* specific region, 1.80% for the core region and 2.11% for the right terminal region. It is clear that there is a need to further screen these genes by increasing the range of *Streptomyces* species analyzed by microarray hybridization. This will reduce the number to a manageable number and will allow prioritization of genes for knockout and detailed phenotypic analysis.

Another approach to the problem of identifying functionally important genes is by the pinpointing of functional groups of such genes that may form a transcriptional unit. Blocks of three or more hypothetical genes that are conserved across all species were identified and are shown in Table 8. It is possible that these groups represent conserved functional groups of genes essential to core functions that make *Streptomyces* different from other bacteria. They are found mostly in the area between the *Streptomyces* terminal regions and the central core region. There are seven groups of conserved hypothetical genes larger than five genes (SCO1407–SCO1413, SCO2362–SCO2370, SCO2911–SCO2919, SCO3846–SCO3854, SCO5536–SCO5543, SCO5762–SCO5767 and SCO6522–6528). It is likely, due to the proximity of various genes around SCO3846–SCO3854, that this complex is involved in cell division, development and

Table 8 Hypothetical genes in *S. coelicolor* conserved as a group in the four species analyzed

Genes (SCO)	Operon structure ^a	Linked function if any ^b
0614, 0616, 0617, 0618	None	–
1317, 1318, 1319, 1320	None	–
1521, 1522, 1523, 1524	Possible operon	Recombination
1634, 1635, 1636	Possible operon	–
1650, 1651, 1652, 1653	Possible operon	Proteosome
1788, 1789, 1790, 1791, 1794, 1795, 1796	Possible operon	Both flanks of rRNA gene homologues
2030, 2031, 2032	Possible operon	–
2124, 2125, 2127, 2129, 2130	Possible operon	Glucose kinase
2268, 2269, 2270	Possible operon	Close to heme oxygenase
2913, 2915, 2916, 2917	None	–
3115, 3117, 3118, 3119	None	–
3150, 3151, 3152, 3153	None	–
3406, 3407, 3408	Possible operon	Penicillin binding protein
3950, 3951, 3952	Possible operon	Oxidoreductase
4028, 4029, 4030	None	–
4801, 4803, 4804, 4805	None	–
5307, 5308, 5309, 5310, 5312	None	–
5600, 5601, 5602, 5603, 5604	Possible operon	Homology to <i>Mycobacterium tuberculosis</i>
5762, 5763, 5764, 5765	Possible operon	DNA helicase
6413, 6415, 6416, 6417, 6419, 6420, 6421, 6422	None	–
6574, 6575, 6576, 6577, 6578, 6579, 6580	Possible operon	Possible DNA binding protein
6671, 6672, 6674, 6675, 6676	Possible operon	–
7070, 7071, 7072	None	–

^a Gene structure from Artemis v7 is compatible with an operon type structure with possible appropriate ribosome binding sites

^b Inside or linked to the conserved genes is a gene(s) of known function

DNA partitioning. The function of the others groups is unknown. Interestingly, none of these gene groups are upregulated shifting from exponential phase to stationary phase or under stress shift as indicated by Karoonuthaisiri et al. (2005).

Conservation of genes involved in secondary metabolism and similar functions

Genes that are involved in secondary metabolism and antibiotic production are widely distributed in the *Streptomyces* and many if not most may have been involved in horizontal transfer. However, there is significant similarity between genes involved in similar pathways and thus significant cross-hybridization may occur between similar metabolic pathways. A large number of genes are also involved in secondary metabolism (165) and polyketide synthesis (102) in the *S. coelicolor* genome. These are grouped together in 23 pathway groups and are displayed in Supplementary Fig. 3. Genes identified as secondary metabolic genes but existing on their own and not in a group of secondary metabolic genes have been eliminated to simplify the analysis leaving only genes involved in these functions with two or more genes together in a group. These include specific pathways producing secondary metabolic products such as melanin, actinorhodin, CDA and Red pathway. Many of the other potential pathways have not been studied in detail and the functions of these genes are unknown. Because of evolutionary similarity, the presence of genes hybridizing to a particular pathway does not mean that the specific pathway is present, but possibly that a related one is may be present. Similarly, a high level of hybridization can mean either a very close relationship between the pathways in the two species or the presence of multiple copies of related pathways. In general terms *S. maritimus* shows the greatest absence of secondary metabolic pathways that are present in *S. coelicolor*. Interesting, *S. cattleya* and *K. aureofaciens* seems to have pathways related to many of the *S. coelicolor* secondary metabolic pathways present in their genomes, although they are phylogenetically more distant than *S. maritimus*. The actinorhodin pathway would seem to be absent from *S. avermitilis* (as expected from the genome data), *S. cattleya* and *S. maritimus* although some related genes do seem to be present in *K. aureofaciens*. The WhiE pathway is

conserved in all species, but with some genes showing a very low level of hybridization in certain cases and these include whiE protein VII and the acyl carrier protein. Genes from the Red pathway show varying levels of hybridization suggesting that distantly related pathways may be present in these species. The CDA pathway is conserved in all four species and in certain cases the genes seem to be over represented suggesting multiple examples of the same type of pathway in *S. cattleya* and *S. maritimus*. The presence of similar pathways at a level of about 50% for *K. aureofaciens* supports the well established idea that horizontal gene transfer of secondary metabolic pathways may have played a significant role in the evolution of the *Streptomyces* and any related genus.

Because the natural environment of *Streptomyces* is the soil, they are thought to play an important role in the recycling of lignocellulose material. However, there is relatively little information on what genes are involved in this process. Interestingly, *melC1* and *melC2*, which encode tyrosinase (monophenol monooxygenase, SCO2700) and its cofactor (SCO2701) (Leu et al. 1992) are conserved across the three *Streptomyces* species and probably also *Kitasatospora* (SCO2700 –0.76, SCO2701 0.08). On the other hand, the duplicate MelD1 (SCO2701) and MelD2 (SCO2700) genes found in *S. coelicolor* are not conserved and are phylogenetically distinct from MelC1 and MelC2 found in other *Streptomyces* (unpublished results). This perhaps represents a divergence of function between this two gene pairs. *S. coelicolor* does not produce a detectable amount of black melanin pigment and these results suggest that these enzymes may be involved in the metabolic conversion of lignocellulose byproducts rather than pigment formation. Evolutionary conservation of these genes to serve this function under particular conditions of induction would make more sense than retention of inducible black pigment formation. Other enzymes with a possible role in the lignocellulose cycle that are conserved across the species are shown in Table 9. These include many oxygenases that may have a role in producing oxygen radicals capable of attacking lignin, genes involved in the sensing and breaking down hydrogen peroxide, cellulose metabolism genes, cellobiose metabolism genes, etc. Those found in the terminal regions may represent gene groups that are not conserved in a syntenous manner and subject to horizontal gene transfer,

Table 9 Genes conserved in the four *Streptomyces* species that are potentially involved in lignocellulose cycling

SCO0333	Dioxygenase
SCO0560	Catalase/Peroxidase
SCO0765	Endoglucanase
SCO1187	Cellulase
SCO1188	Cellulose binding protein
SCO1338	Monooxygenase
SCO1451	Endoglucanase
SCO1923	Dioxygenase
SCO2016	Monooxygenase
SCO2267	Heme oxygenase
SCO2700	Tyrosinase (monophenol monooxygenase)
SCO2701	Tyrosinase cofactor
SCO2783	Monooxygenase
SCO2798	Cellobiose hydrolase
SCO2838	Endoglucanase
SCO3172	Monooxygenase
SCO3236	Oxygenase
SCO4416	Monooxygenase
SCO4870	Monooxygenase
SCO5033	Hydrogen peroxide sensing regulator
SCO5293	Oxygenase
SCO5390	Alkanal monooxygenase
SCO5773	Monooxygenase
SCO6545	Cellulase
SCO7223	Monooxygenase
SCO7637	Endoglucanase

Note that the oxygeneases included as possible enzymes that make be able to attack lignin are all unclassified yet as to their real function. The core region is in bold

while those within the core and intermediate regions may be part of the basic group of genes essential to *Streptomyces* in the soil environment. Lignocellulose degradation is a difficult topic to study in the Actinomycetales and therefore these candidate genes may help to solve some of the problems associated with this.

Conclusions

This study confirms that within the *Streptomyces* analyzed here there is conservation of a core set of genes in the middle of the linear *S. coelicolor*/

S. avermitilis chromosome structure. This is associated with a much higher diversity of gene in the terminal regions of the linear chromosome. Linking these regions are two intermediate regions where there seems to be conservation of genus specific genes and gene clusters. This study also identifies candidate genes that may be possibly involved in terminal replication and other myceliate growth related functions based on a classification of genes into conserved and none conserved groups. This study also provides insights into which genes in *Streptomyces* play a more significant role in the biochemical network of *S. coelicolor*, *Streptomyces* and myceliate Actinobacteria in general. Finally, the degree of gene conserved detected between the four species implies that that genome model of *S. coelicolor* may extent well beyond the borders of the *Streptomyces*. It includes at least one *Kitasatospora* species; furthermore, a similar structure by microarray analysis has been found for *Saccharomonospora viridis* and *Streptosporangium roseum*, but not *Streptomyces rimosus* ATCC10970 (unpublished data). Thus, the microarray approach to genome content analysis and exploration of genome evolution may be fairly widely applicable in the various Actinomycete genus close to *Streptomyces* that undergo complex morphogenesis.

Acknowledgements We would like to thank Cammy M. Kao and David W. Weaver for access to the Stanford microarrays, for help setting up the system and for their many helpful suggestions. Special thanks goes to Carton Chen and the rest of the *Streptomyces* research team at Yang-Ming without whose help and support this work would not have been possible. Financial support is acknowledged from National Yang-Ming University, Taiwan and the National Science Council, Taiwan (93-2320-B-010-041 and 94-2320-B-010-037).

References

- Akman L, Aksoy S (2001) A novel application of gene arrays: *Escherichia coli* array provides insights into the biology of the obligate endosymbiote of Tsetse flies. Proc Natl Acad Sci USA 98:7546–7551
- Akman L, Rio RV, Beard CB, Aksoy S (2001) Genome size determination, and coding capacity of *Sodalis glossinidius*, an enteric symbiont of Tsetse flies, as revealed by hybridization to *Escherichia coli* gene arrays. J Bacteriol 183:4517–4525
- Bao K, Cohen SN (2001) Terminal proteins essential for the replication of linear plasmids, and chromosomes in *Streptomyces*. Genes Dev 15:1518–1527

- Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 284:1520–1523
- Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A32. *Nature* 417:141–147
- Chan K, Baker S, Kim CC, Detweiler CS, Dougan G., Falkow S (2003) Genomic comparison of *Salmonella bongori* by use of an *S. enterica* serovar Typhimurium DNA microarray. *J Bacteriol* 185:553–563
- Chen CW, Huang CH, Lee HH, Tsai HH, Kirby R (2002) Once the circle has been broken: dynamics, and evolution of *Streptomyces* chromosomes. *Trends Genet* 18:522–529
- Cho J-C, Tiedje JM (2001) Bacterial species determination from DNA-DNA hybridization by using genome fragments, and DNA microarrays. *Appl Environ Microbiol* 2001 67:3677–3682
- Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou FX, Fourrier C, Guerineau M, Decaris B, Barbe V, Pernodet JL, Leblond P (2006) Evolution of the terminal regions of the *Streptomyces* linear chromosome. *Mol Biol Evol* 23:2361–2369
- Choulet F, Gallois A, Aigle B, Mangenot S, Gerbaud C, Truong C, Francou FX, Borges F, Fourrier C, Guerineau M, Decaris B, Barbe V, Pernodet JL, Leblond P (2006) Intraspecific variability of the terminal inverted repeats of the linear chromosome of *Streptomyces ambofaciens*. *J Bacteriol* 188:6599–6610
- Dary A, Martin P, Wenner T, Decaris B, Leblond P (2000) DNA rearrangements at the extremities of the *Streptomyces ambofaciens* linear chromosome: evidence for developmental control. *Biochimie* 82:29–34
- Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al-Ghusein H, Barrell BG, Parkhill J, Stoker NG, Karlyshev AV, Butcher PD, Wren BW. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low cost microarray reveals extensive genetic diversity. *Genome Res* 11:1706–1715
- Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ (2002) Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic, and pandemic disease. *Proc Natl Acad Sci USA* 99:1556–1561
- Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis, and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868
- Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM (2001) Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains, and the toxic shock syndrome epidemic. *Proc Natl Acad Sci USA* 98:8821–8826
- Gill RT, Wildt S, Yang YT, Ziesman S, Stephanopoulos G (2002) Genome-wide screening for trait conferring genes using DNA microarrays. *Proc Natl Acad Sci USA* 99:7033–7038
- Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, Schroeder M, Brown PO, Botstein D, Sherlock G. (2003) Microarray database: data access, and quality assessment tools. *Nucleic Acids Res* 31:94–96
- Goodner BW, Markelz BP, Flanagan MC, Crowell CB Jr, Racette JL, Schilling BA, Halfon LM, Mellors JS, Gabowski G. (1999) Combined genetic, and physical map of the complex genome of *Agrobacterium tumefaciens*. *J Bacteriol* 181:5160–5166
- Huang J, Lih CJ, Pan KH, Cohen SN (2001) Global analysis of growth phase responsive gene expression, and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays. *Genes Dev* 15:3183–3192
- Huang CH, Lin YS, Yang YL, Huang SW, Chen CW (1998) The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. *Mol Microbiol* 28:905–916
- Huang WM, Robertson M, Aron J, Casjens S. (2004) Telomere exchange between linear replicons of *Borrelia burgdorferi*. *J Bacteriol* 186:4134–4141
- Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S. (2003) Complete genome sequence comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* 21:526–531
- Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, Peek RM Jr (2001) *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc Natl Acad Sci USA* 98:14625–14630
- Kamiya H, Ishiguro C, Harashima H (2004) Increased A:T → C:G mutations in the mutT strain upon 8-hydroxy-dGTP treatment: direct evidence for MutT involvement in the prevention of mutations by oxidized dGTP. *J Biochem Tokyo* 136:359–362
- Karoonuthaisiri N, Weaver D, Huang J, Cohen SN, Kao C (2005) Regional organization of gene expression in *Streptomyces coelicolor*. *Gene* 353:53–66
- Kobayashi T, Takao M, Oikawa A, Yasui A (1989) Molecular characterization of a gene encoding a photolyase from *Streptomyces griseus*. *Nucleic Acids Res* 17:4731–4744
- Leonard EE 2nd, Takata T, Blaser MJ, Falkow S, Tompkins LS, Gaynor EC (2003) Use of an open-reading frame-specific *Campylobacter jejuni* DNA microarray as a new genotyping tool for studying epidemiologically related isolates. *J Infect Dis* 187:691–694
- Leu WM, Chen LY, Liaw LL, Lee YH (1992) Secretion of the *Streptomyces* tyrosinase is mediated through its trans-activator protein MelC1. *J Biol Chem* 267:20108–20113
- Lin YS, Chen CW (1997) Instability of artificially circularized chromosomes of *Streptomyces lividans*. *Mol Microbiol* 26:709–719
- Lin YS, Kieser HM, Hopwood DA, Chen CW (1993) The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol Microbiol* 10:923–933

- Lum AM, Huang J-Q, Hutchinson CR, Kao CM (2004) Reverse engineering of industrial pharmaceutical-producing actinomycete strains using DNA microarrays. *Metab Eng* 6:186–196
- Murray AE, Lies D, Li G, Nealson K, Zhou J, Tiedje JM (2001) DNA/DNA hybridization to microarrays reveals gene specific differences between closely related microbial genomes. *Proc Natl Acad Sci USA* 98:9853–9858
- Pospiech A, Neumann B (1995) A versatile quick-prep of genomic DNA from Gram-positive bacteria. *Trends Genet* 11:217–218
- Porwollik S, Wong RM, McClell M (2002) Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. *Proc Natl Acad Sci USA* 99:8956–8961
- Rajashekara G, Glasner JD, Glover DA, Splitter GA (2004) Comparative whole-genome hybridization reveals genomic islands in *Brucella* species. *J Bacteriol* 186:5040–5051
- Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci USA* 97:14668–14673
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acid Res* 24:4876–4882
- Vinciotti V, Khanin R, D’Alimonte D, Liu X, Cattini N, Hotchkiss G, Bucca G, de Jesus O, Rasaiyaah J, Smith CP, Kellam P, Wit E (2005) An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics* 15:492–501
- Wang SJ, Chang HM, Lin YS, Huang CH, Chen CW (1999) *Streptomyces* genomes: circular genetic maps from the linear chromosomes. *Microbiology* 145:2209–2220
- Williams ST, Goodfellow M, Alderson G, Wellington EM, Sneath PH, Sackin MJ (1983a) Numerical classification of *Streptomyces* related genera. *J Gen Microbiol* 129:1743–813
- Williams ST, Goodfellow M, Wellington EM, Vickers JC, Alderson G, Sneath PH, Sackin MJ, Mortimer AM (1983b) A probability matrix for identification of some *Streptomyces*. *J Gen Microbiol* 129:1815–1830
- Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees the tree of life. *Trends Genet* 18:472–479
- Yang CC, Huang CH, Li CY, Tsay YG, Lee SC, Chen CW (2002) The terminal proteins of linear *Streptomyces* chromosomes plasmids: a novel class of replication priming proteins. *Mol Microbiol* 43:297–305
- Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WP, del Cardayre SB (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415:644–646